情感分析的基础实验

李牧之

2025 年 7 月 29 日

1 Introduction

基于 python 和 IMDB 电影评论数据集，运用词袋 (BOW)、词频 (TFIDF)、深度学习方法进行基础实验，并对模型性能进行简单对比评估

2 Dataset

2.1 数据集信息

IMDB 电影评论数据集，包含 5 万条英语原始评论，每条评论含有真实的情感倾向标签，并且被平均分为了积极与消极两部分。
下载链接：

`https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews`

2.2 数据处理

2.2.1 数据清洗 Data Cleaning

对于数据集中的数据，我们仅考虑其中的词汇，因此要去除特殊符号、数字等；同时将大小写归一化。

2.2.2 停用词过滤 Stop Words Removal

停用词（Stop Words）是指在自然语言文本中出现频率很高但对文本意义贡献很小的词，将这些词语过滤掉。

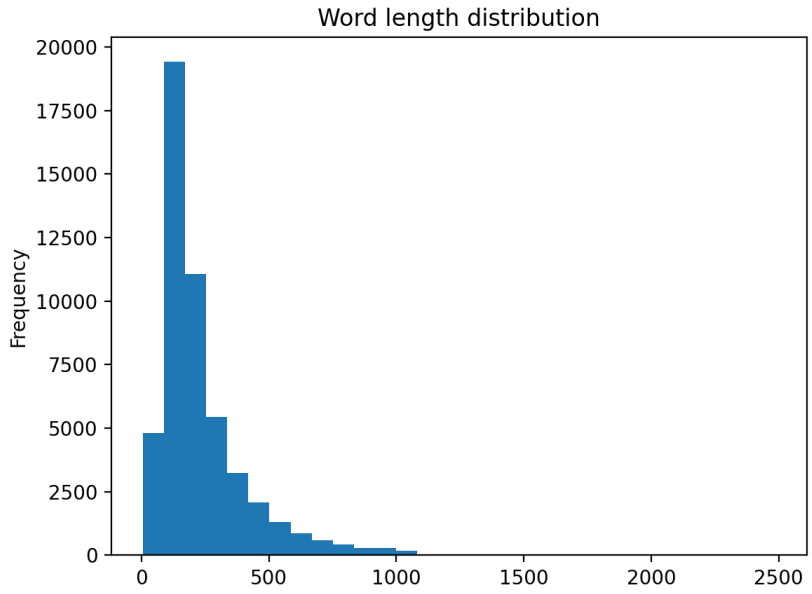
2.2.3 词干提取 Stemming

将一个词语的不同形式还原成词干，通过去除前后缀来进行，将具有相同基本意义的词语归一化。

2.2.4 深度学习方法所需要的进一步数据处理

深度学习方法不需要进行停用词过滤以及词干提取。在数据清理之后，进行词汇表的创建，每一个词语赋予独特的索引值，将评论文本转换为一串整数。大约共有 92K 的单词，其中前 10K 的单词就可

以覆盖文中约 95% 的单词。所以仅考虑前 10K 的单词，并将单一评论长度限制为 500 词，短于 500 词用特殊索引填充，长于 500 词则只考虑前 500 词进行截取。



2.3 数据集划分

40K 条数据作为训练集，10K 条数据作为测试集，随机划分。

3 Experiments

3.1 词袋 Bags of Words

使用 sklearn 库中的 CountVectorize 实现一个词、一个词 + 两个词、一个词 + 两个词 + 三个词的三种词袋。仅通过训练集中的评论来构建词汇表，且同一词汇表将被应用于测试集。分别使用逻辑回归 (Logistic Regression)、线性支持向量机 (LSVM)、朴素贝叶斯 (NB) 进行机器学习。

表 1: 基于词袋实验准确率结果

Methods	Unigrams	Uni+Bigrams	Uni+Bi+Trigrams
Logistic Regression	88%	90%	90%
LSVM	86%	90%	89%
NB	85%	88%	89%

3.2 词频 TFIDF

应用 TDIDF 分数，在多个文本同时存在的情况下，对于一个词语在某一给定文本中的重要性进行评估。分别使用逻辑回归 (Logistic Regression)、线性支持向量机 (LSVM)、朴素贝叶斯 (NB) 进行机

器学习。

表 2: 基于词频实验准确率结果

Methods	Unigrams	Uni+Bigrams	Uni+Bi+Trigrams
Logistic Regression	89%	89%	88%
LSVM	89%	90%	90%
NB	86%	89%	89%

3.3 深度学习方法 Deep Learning

使用 2.2.4进一步处理后的数据, 应用基于 Tensorflow 的多层感知机 (MLP)、循环神经网络 (RNN)、长短时神经网络 (LSTM)、1D 卷积神经网络 (1D CNN) 进行深度学习。

表 3: 深度学习实验准确率结果

Methods	Accuracies
MLP	87.0%
RNN	84.0%
LSTM	88.9%
1D CNN	89.7%

4 Summary

表 4: 实验准确率结果汇总

Methods	Accuracies
Logistic Regression with BOW	90.0%
LSVM with BOW	90.0%
NB with BOW	89.0%
Logistic Regression with TFIDF	89.0%
LSVM with TFIDF	90.0%
NB with TFIDF	89.0%
MLP	87.0%
RNN	84.0%
LSTM	88.9%
1D CNN	89.7%

5 More Information

5.1 References

1.https://dropsofai.com/sentiment-analysis-with-python-bag-of-words/

2.https://dropsofai.com/sentiment-analysis-with-python-tfidf-features/

3.https://dropsofai.com/sentiment-classification-with-deep-learning-rnn-lstm-and-cnn/

5.2 Github Link

https://github.com/GabrielMu2006/Sentiment-Analysis